

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-372987
 (43)Date of publication of application : 26.12.2002

(51)Int.CI. G10L 15/14
 G10L 15/06

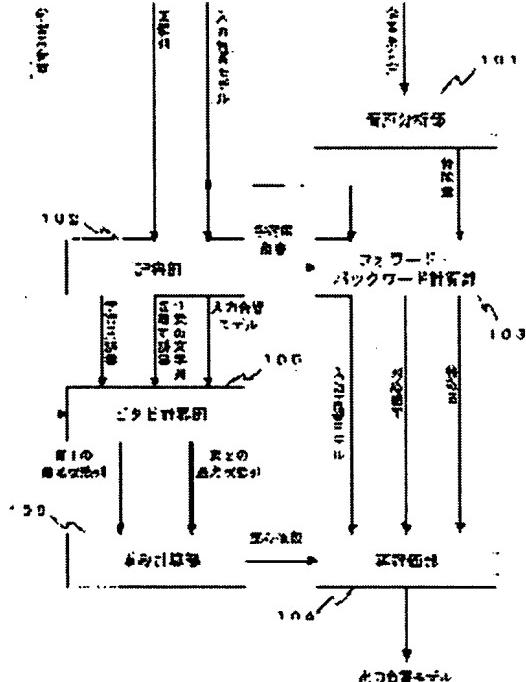
(21)Application number : 2001-179125 (71)Applicant : NEC CORP
 (22)Date of filing : 13.06.2001 (72)Inventor : TAKANO MASARU

(54) ACOUSTIC MODEL LEARNING DEVICE, ACOUSTIC MODEL LEARNING METHOD, AND PROGRAM FOR THE SAME

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an acoustic model learning device, an acoustic model learning method and a program for the same for extracting only a voice sample useful for generating an acoustic model out of observed voice samples to generate a highly reliable acoustic model.

SOLUTION: The reevaluation section 104 calculates a statistical quantity based on the feature amount of the voice for learning extracted by a voice analyzing section 104 and correspondence probability calculated by a forward/ backward calculating section 103 and a weight coefficient Rt calculated by a weight calculating section 106 to re-estimate the acoustic model and output the output acoustic model.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2002-372987
(P2002-372987A)

(43) 公開日 平成14年12月26日 (2002. 12. 26)

(51) Int.Cl.⁷
G 1 0 L 15/14
15/06

識別記号

F I
G 1 0 L 3/00

テマコト^{*} (参考)
5 3 5 Z 5 D 0 1 5
5 2 1 F

審査請求 未請求 請求項の数18 O.L. (全 17 頁)

(21) 出願番号 特願2001-179125(P2001-179125)

(22) 出願日 平成13年6月13日 (2001. 6. 13)

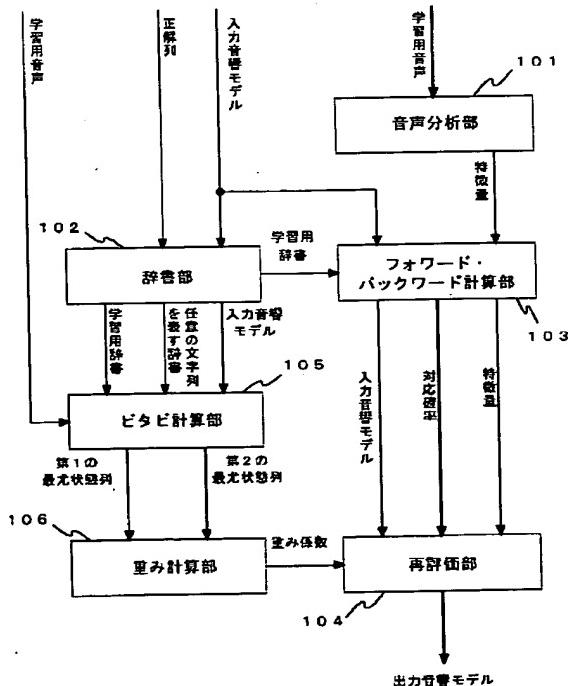
(71) 出願人 000004237
日本電気株式会社
東京都港区芝五丁目7番1号
(72) 発明者 高野 優
東京都港区芝五丁目7番1号 日本電気株
式会社内
(74) 代理人 100084250
弁理士 丸山 隆夫
F ターム(参考) 5D015 GG00 HH00

(54) 【発明の名称】 音響モデル学習装置、音響モデル学習方法、およびそのプログラム

(57) 【要約】

【課題】 観測された音声サンプルのうち音響モデルの作成に有用なものだけを抽出し、信頼性の高い音響モデルを作成する音響モデル学習装置、音響モデル学習方法、およびプログラムを提供する。

【解決手段】 再評価部104は、音声分析部101により抽出された学習用音声の特徴量と、フォワード・バックワード計算部103により算出された対応確率と、重み計算部106により算出された重み係数R_tと、に基づいて統計量を算出し、音響モデルの再推定を行い、出力音響モデルを出力する。



【特許請求の範囲】

【請求項1】 入力される学習用音声からフレームごとに特徴量を抽出する音声分析手段と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、前記所定の音声のフレームごとに分割された前記所定の音声の断片を状態として表現し、該状態を構成単位とする入力音響モデルと、前記学習用音声の内容を示す文字列情報である正解列と、に基づいて、前記入力音響モデルにおける前記状態に前記正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成手段と、該辞書生成手段により生成された学習用辞書を参照し、前記学習用音声の特徴量と前記入力音響モデルにおける状態との対応確率を前記学習用音声のフレームごとに算出する対応確率算出手段と、所定の文字列を用いて、前記入力音響モデルにより表現される前記状態あるいは複数の前記状態からなる状態列を、前記学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成手段と、該最尤状態列生成手段により生成された所定の最尤状態列に基づいて、前記対応確率に重み付けする際に付加する係数である重み係数を、前記学習用音声のフレームごとに算出する重み計算手段と、前記対応確率算出手段により算出された対応確率と、前記重み計算手段により算出された重み係数と、前記音声分析手段により算出された特徴量と、に基づいて統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価手段と、を有することを特徴とする音響モデル学習装置。

【請求項2】 前記再評価手段は、

前記学習用音声のフレームごとの前記対応確率に、前記重み係数を乗算し、前記学習用音声のフレームごとの対応確率に重み付けを行い、該重み付けされた対応確率を用いて前記統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、前記出力音響モデルを作成することを特徴とする請求項1記載の音響モデル学習装置。

【請求項3】 前記重み計算手段は、前記最尤状態列生成手段により、前記学習用辞書を用いて生成された最尤状態列を第1の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第2の最尤状態列とした場合、前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、該比較に基づいて、前記学習用音声のフレームごとに前記重み係数を算出することを特徴とする請求項1または2記載の音響モデル学習装置。

【請求項4】 前記重み計算手段は、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、前記割り当てら

れた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を1とし、互いに異なるフレームでは前記重み係数を1より小さな値として算出することを特徴とする請求項3記載の音響モデル学習装置。

【請求項5】 前記重み計算手段は、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を1とし、互いに異なるフレームでは前記重み係数を1より大きな値として算出することを特徴とする請求項3記載の音響モデル学習装置。

【請求項6】 前記重み計算手段は、

前記割り当てられた状態ごとに、算出した前記重み係数の和をそれぞれ算出し、該算出した重み係数の和が、それぞれ等しい値となるように前記算出した重み係数を設定することを特徴とする請求項1からうのいずれか1項に記載の音響モデル学習装置。

【請求項7】 入力される学習用音声からフレームごとに特徴量を抽出する音声分析工程と、

所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、前記所定の音声におけるフレームごとに分割された前記所定の音声の断片を状態として表現し、該状態を構成単位とする入力音響モデルと、前記学習用音声の内容を示す文字列情報である正解列と、に基づいて、前記入力音響モデルにおける前記状態に前記正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成工程と、

該辞書生成工程により生成された学習用辞書を参照し、前記学習用音声の特徴量と前記入力音響モデルにおける状態との対応確率を前記学習用音声のフレームごとに算出する対応確率算出工程と、

所定の文字列を用いて、前記入力音響モデルにより表現される前記状態あるいは複数の前記状態からなる状態列を、前記学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成工程と、該最尤状態列生成工程により生成された所定の最尤状態列に基づいて、前記対応確率に重み付けする際に付加する係数である重み係数を、前記学習用音声のフレームごとに算出する重み計算工程と、

前記対応確率算出工程により算出された対応確率と、前記重み計算工程により算出された重み係数と、前記音声分析工程により算出された特徴量と、に基づいて統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価工程と、

を有することを特徴とする音響モデル学習方法。

【請求項8】 前記再評価工程は、

前記学習用音声のフレームごとの前記対応確率に、前記重み係数を乗算し、前記学習用音声のフレームごとの対応確率に重み付けを行い、該重み付けされた対応確率を

用いて前記統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、前記出力音響モデルを作成することを特徴とする請求項7記載の音響モデル学習方法。

【請求項9】 前記重み計算工程は、

前記最尤状態列生成工程により、前記学習用辞書を用いて生成された最尤状態列を第1の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第2の最尤状態列とした場合、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、該比較に基づいて、前記学習用音声のフレームごとに前記重み係数を算出することを特徴とする請求項7または8記載の音響モデル学習方法。

【請求項10】 前記重み計算工程は、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を1とし、互いに異なるフレームでは前記重み係数を1より小さな値として算出することを特徴とする請求項9記載の音響モデル学習方法。

【請求項11】 前記重み計算工程は、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を1とし、互いに異なるフレームでは前記重み係数を1より大きな値として算出することを特徴とする請求項9記載の音響モデル学習方法。

【請求項12】 前記重み計算工程は、

前記割り当てられた状態ごとに、算出した前記重み係数の和をそれぞれ算出し、該算出した重み係数の和が、それぞれ等しい値となるように前記算出した重み係数を設定することを特徴とする請求項7から11のいずれか1項に記載の音響モデル学習方法。

【請求項13】 入力される学習用音声からフレームごとに特徴量を抽出する音声分析処理と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、前記所定の音声におけるフレームごとに分割された前記所定の音声の断片を状態として表現し、該状態を構成単位とする入力音響モデルと、前記学習用音声の内容を示す文字列情報である正解列と、に基づいて、前記入力音響モデルにおける前記状態に前記正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成処理と、

該辞書生成処理により生成された学習用辞書を参照し、前記学習用音声の特徴量と前記入力音響モデルにおける状態との対応確率を前記学習用音声のフレームごとに算出する対応確率算出処理と、

所定の文字列を用いて、前記入力音響モデルにより表現される前記状態あるいは複数の前記状態からなる状態列

を、前記学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成処理と、該最尤状態列生成処理により生成された所定の最尤状態列に基づいて、前記対応確率に重み付けする際に付加する係数である重み係数を、前記学習用音声のフレームごとに算出する重み計算処理と、

前記対応確率算出処理により算出された対応確率と、前記重み計算処理により算出された重み係数と、前記音声分析処理により算出された特徴量と、に基づいて統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価処理と、

を実行させるためのプログラム。

【請求項14】 前記再評価処理は、

前記学習用音声のフレームごとの前記対応確率に、前記重み係数を乗算し、前記学習用音声のフレームごとの対応確率に重み付けを行い、該重み付けされた対応確率を用いて前記統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、前記出力音響モデルを作成することを特徴とする請求項13記載のプログラム。

【請求項15】 前記重み計算処理は、

前記最尤状態列生成処理により、前記学習用辞書を用いて生成された最尤状態列を第1の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第2の最尤状態列とした場合、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、該比較に基づいて、前記学習用音声のフレームごとに前記重み係数を算出することを特徴とする請求項13または14記載のプログラム。

【請求項16】 前記重み計算処理は、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を1とし、互いに異なるフレームでは前記重み係数を1より小さな値として算出することを特徴とする請求項15記載のプログラム。

【請求項17】 前記重み計算処理は、

前記学習用音声のフレームごとに、前記第1の最尤状態列と前記第2の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を1とし、互いに異なるフレームでは前記重み係数を1より大きな値として算出することを特徴とする請求項15記載の音響モデル学習装置。

【請求項18】 前記重み計算処理は、

前記割り当てられた状態ごとに、算出した前記重み係数の和をそれぞれ算出し、該算出した重み係数の和が、それぞれ等しい値となるように前記算出した重み係数を設

定することを特徴とする請求項13から17のいずれか1項に記載の音響モデル学習装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、音響モデル学習装置、音響モデル学習方法、およびそのプログラムに関し、特に、音声サンプルの特性に応じて音声サンプルに重み付けを行い、信頼性の高い音響モデルを作成する音響モデル学習装置、音響モデル学習方法、およびそのプログラムに関する。

【0002】

【従来の技術】音響モデル学習装置は、実際の音声を用いて、音声認識に使用される音響モデルを学習することが多い。一般に、学習される音響モデルとして、Hidden Markov Model (隠れマルコフモデル、以下、HMMとする) が用いられる。また、HMMにおける状態を表す確率分布としては、連続混合分布が用いられる場合が多い。また、多くの場合、HMMの学習には、フォワード・バックワード法が用いられる。上

$$\alpha(1, i) = b(i, O_1) \dots \quad (\text{式1.1})$$

$$\alpha(t+1, j) = [\sum_{i=1}^N \alpha(t, i) a_{ij}] b(j, O_{t+1}) \dots \quad (\text{式1.2})$$

【0007】なお、フォワード確率 $\alpha(t, i)$ は、特徴量 O_t を観測し、状態 S_i にある確率を示す。同様に、フォワード確率 $\alpha(1, i)$ は、特徴量 O_1 を観測し、状態 S_i にある確率、フォワード確率 $\alpha(t+1, j)$ は、特徴量 O_{t+1} を観測し、状態 S_j にある確率を示す。

【0008】また、状態遷移確率 a_{ij} は、状態 S_i から状態 S_j へ遷移する確率を表す。観測確率 $b(i,$

$$\beta(T, i) = 1 \dots \quad (\text{式2.1})$$

$$\beta(t, i) = \sum_{j=1}^N a_{ij} b(j, O_{t+1}) \beta(t+1, j) \dots \quad (\text{式2.2})$$

【0011】なお、バックワード確率 $\beta(t, i)$ は、フレーム t において状態 S_i にあり、以後フレーム $(t+1)$ において特徴量 O_{t+1} を観測する確率を示す。フレーム T は、最終状態におけるフレームを表す。

【0012】また、フォワード・バックワード法における

$$\gamma(t, j, k) = \frac{\alpha(t, j) \beta(t, j)}{N} c_{jk} N(O_t, \mu_{jk}, U_{jk})$$

記のようなHMMによる音響モデルのパラメータの推定について記載されている文献としては、Lawrence Labiner, Bing-Hwang Juang 「Fundamentals of Speech Recognition 1993 p. 333~p. 389」(以下、従来例1) があった。

【0003】従来例1では、HMMに用いられる連続混合確率分布を構成する複数の確率分布それぞれに、連続混合確率分布における混合比を示す混合重みを付加していた。

【0004】以下、フォワード・バックワード法を用いたHMMにおけるパラメータの計算方法について説明する。

【0005】時刻(フレーム) t ごとの特徴量を O_t (t は 1 以上 T 以下の整数) とすると、フォワード・バックワード法におけるフォワード確率 α は、以下に示す(式1.1)および(式1.2)により示される。

【0006】

【数1】

$$\dots \quad (\text{式1.1})$$

O_1) は、状態 S_i に遷移する際に、フレーム t における特徴量 O_t が観測される確率を示す。

【0009】また、フォワード・バックワード法におけるバックワード確率 β は、以下に示す(式2.1)および(式2.2)により示される。

【0010】

【数2】

$$\dots \quad (\text{式2.1})$$

る対応確率 γ は、フォワード確率 α とバックワード確率 β とに基づいて、計算される。対応確率 γ は、以下に示す(式3.1)により示される。

【0013】

【数3】

クトル $\mu(t, j, k)$ 、および共分散行列 $U(j, k)$ の各平均は、以下に示す(式4.1)、(式4.2)、および(式4.3)により計算される。

【0016】
【数4】

$$\bar{c}_{jk} = \frac{\sum_{t=1}^T \gamma(t, j, k)}{\sum_{t=1}^T \sum_{k=1}^n \gamma(t, j, k)} \quad \dots \quad (\text{式4.1})$$

$$\bar{\mu}(t, j, k) = \frac{\sum_{t=1}^T \gamma(t, j, k) \cdot O_t}{\sum_{t=1}^T \gamma(t, j, k)} \quad \dots \quad (\text{式4.2})$$

$$\bar{U}(j, k) = \frac{\sum_{t=1}^T \gamma(t, j, k) \cdot (O_t - \bar{\mu}_{jk})(O_t - \bar{\mu}_{jk})'}{\sum_{t=1}^T \gamma(t, j, k)} \quad \dots \quad (\text{式4.3})$$

【0017】なお、混合重み c_{jk} は、HMMにおける状態 S_j のk番目の混合分布要素に対する混合重みである。また、平均ベクトル $\mu(t, j, k)$ は、HMMにおける状態 S_j のk番目の混合分布要素の平均ベクトルである。また、共分散行列 $U(j, k)$ は、HMMにおける状態 S_j のk番目の混合分布要素の共分散行列である。また、 V_k は、文字列Vにおける所定の文字を示す。また、 $(O_t - \mu_{jk})'$ は、ベクトル $(O_t - \mu_{jk})$ の対置ベクトルを表す。

【0018】また、特開平5-232989号公報が開示するところの音響モデルの話者適応化法(以下、従来例2)では、HMMに用いられる連続混合確率分布を構成する複数の確率分布それぞれの混合比を決める重み係数だけを再推定していた。

【0019】また、特開平10-11086号公報が開示するところの隠れマルコフモデルの計算方式(以下、従来例3)には、フォワードバックワード法を用いたHMMの計算方式が記載されていた。

【0020】

【発明が解決しようとする課題】一般に、信頼性の高い確率モデルの学習には、大量の音声データが必要となる。特に、不特定話者用の音響モデルには、話者の個人差による音声の変動を吸収する必要がある。従って、不特定話者用の音響モデルには、話者の発声による音声データが多数必要となる。しかしながら、大量の音声サンプルを収集する際には、話者の誤発声あるいは低品質の

音声が混入する可能性がある。

【0021】さらに、確率モデル(音響モデル)の推定を行う場合に、以下に示すような問題が生じてしまう。通常、音声データを収集する際、話者の自然な発声による音声データを得る必要がある。従って、音声データとして収集される話者の発声内容は、実際に存在する単語が用いられる。また、実際に存在する単語を構成する音素(文字)の分布には必然的に偏りが生じる。例えば、日本語の場合は、母音、特に「あ」の出現頻度が非常に高い。確率モデルを推定する場合、確率分布を推定するサンプル数によって確率分布の信頼性に格差が生じてしまう。従って、単語を構成する音素を音響モデルを構築する音声データとして用いる場合、音素の出現頻度の偏りを修正する必要がある。

【0022】本発明は、上記問題点に鑑みてなされたものであり、従来例1、従来例2、および従来例2と従来例3とを組み合わせたものにおいてHMMの各混合分布要素に付加されている重みに加え、収集した音声サンプルの特性に応じて設定された重み係数を、音声サンプルの各フレームにさらに付加することによって、特定の音声サンプルあるいは音声サンプルの特定部分を音響モデルの学習の際に増幅あるいは除去し、音声サンプルを構成する音素の出現頻度の偏りを修正し、信頼性の高い音響モデルを提供する音響モデル学習装置を提供することを目的とする。

【0023】

【課題を解決するための手段】かかる目的を達成するため、請求項1記載の発明は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析手段と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとに分割された所定の音声の断片を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成手段と、辞書生成手段により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出手段と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成手段と、最尤状態列生成手段により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算手段と、対応確率算出手段により算出された対応確率と、重み計算手段により算出された重み係数と、音声分析手段により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価手段と、を有することを特徴とする。

【0024】また、請求項2記載の発明によれば、請求項1記載の音響モデル学習装置において、再評価手段は、学習用音声のフレームごとの対応確率に、重み係数を乗算し、学習用音声のフレームごとの対応確率に重み付けを行い、重み付けされた対応確率を用いて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成することを特徴とする。

【0025】また、請求項3記載の発明によれば、請求項1または2記載の音響モデル学習装置において、重み計算手段は、最尤状態列生成手段により、学習用辞書を用いて生成された最尤状態列を第1の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第2の最尤状態列とした場合、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、比較に基づいて、学習用音声のフレームごとに重み係数を算出することを特徴とする。

【0026】また、請求項4記載の発明によれば、請求項3記載の音響モデル学習装置において、重み計算手段は、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を1とし、互いに異なるフレームでは重み係数を1より小さな値として算出することを特徴とする。

【0027】また、請求項5記載の発明によれば、請求項3記載の音響モデル学習装置において、重み計算手段は、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を1とし、互いに異なるフレームでは重み係数を1より大きな値として算出することを特徴とする。

【0028】また、請求項6記載の発明によれば、請求項1から5のいずれか1項に記載の音響モデル学習装置において、重み計算手段は、割り当てられた状態ごとに、算出した重み係数の和をそれぞれ算出し、算出した重み係数の和が、それぞれ等しい値となるように算出した重み係数を設定することを特徴とする。

【0029】また、請求項7記載の発明は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析工程と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとに分割された所定の音声の断片を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成工程と、辞書生成工程により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出工程と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成工程と、最尤状態列生成工程により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算工程と、対応確率算出工程により算出された対応確率と、重み計算工程により算出された重み係数と、音声分析工程により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価工程と、を有することを特徴とする音響モデル学習方法。

【0030】また、請求項8記載の発明によれば、請求項7記載の音響モデル学習方法において、再評価工程は、学習用音声のフレームごとの対応確率に、重み係数を乗算し、学習用音声のフレームごとの対応確率に重み付けを行い、重み付けされた対応確率を用いて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成することを特徴とする。

【0031】また、請求項9記載の発明によれば、請求項7または8記載の音響モデル学習方法において、重み計算工程は、最尤状態列生成工程により、学習用辞書を

用いて生成された最尤状態列を第1の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第2の最尤状態列とした場合、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、比較に基づいて、学習用音声のフレームごとに重み係数を算出することを特徴とする。

【0032】また、請求項10記載の発明によれば、請求項9記載の音響モデル学習方法において、重み計算工程は、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を1とし、互いに異なるフレームでは重み係数を1より小さな値として算出することを特徴とする。

【0033】また、請求項11記載の発明によれば、請求項9記載の音響モデル学習方法において、重み計算工程は、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を1とし、互いに異なるフレームでは重み係数を1より大きな値として算出することを特徴とする。

【0034】また、請求項12記載の発明によれば、請求項7から11のいずれか1項に記載の音響モデル学習方法において、重み計算工程は、割り当てられた状態ごとに、算出した重み係数の和をそれぞれ算出し、算出した重み係数の和が、それぞれ等しい値となるように算出した重み係数を設定することを特徴とする。

【0035】また、請求項13記載の発明は、入力される学習用音声からフレームごとに特微量を抽出する音声分析処理と、所定の音声からフレームごとに抽出された特微量を示す確率分布を用いて、所定の音声におけるフレームごとに分割された所定の音声の断片を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報を正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てる状態列の情報である学習用辞書を生成する辞書生成処理と、辞書生成処理により生成された学習用辞書を参照し、学習用音声の特微量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出処理と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当てる、所定の最尤状態列を生成する最尤状態列生成処理と、最尤状態列生成処理により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算処理と、対応確率算出処理により算出された対応確率と、重み計算処理により算出された重み係数と、音声分析処理により算出された特微量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを

作成する再評価処理と、を実行させることを特徴とする。

【0036】また、請求項14記載の発明によれば、請求項13記載のプログラムにおいて、再評価処理は、学習用音声のフレームごとの対応確率に、重み係数を乗算し、学習用音声のフレームごとの対応確率に重み付けを行い、重み付けされた対応確率を用いて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成することを特徴とする。

【0037】また、請求項15記載の発明によれば、請求項13または14記載のプログラムにおいて、重み計算処理は、最尤状態列生成処理により、学習用辞書を用いて生成された最尤状態列を第1の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第2の最尤状態列とした場合、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、比較に基づいて、学習用音声のフレームごとに重み係数を算出することを特徴とする。

【0038】また、請求項16記載の発明によれば、請求項15記載のプログラムにおいて、重み計算処理は、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を1とし、互いに異なるフレームでは重み係数を1より小さな値として算出することを特徴とする。

【0039】また、請求項17記載の発明によれば、請求項15記載の音響モデル学習装置において、重み計算処理は、学習用音声のフレームごとに、第1の最尤状態列と第2の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を1とし、互いに異なるフレームでは重み係数を1より大きな値として算出することを特徴とする。

【0040】また、請求項18記載の発明によれば、請求項13から17のいずれか1項に記載の音響モデル学習装置において、重み計算処理は、割り当てられた状態ごとに、算出した重み係数の和をそれぞれ算出し、算出した重み係数の和が、それぞれ等しい値となるように算出した重み係数を設定することを特徴とする。

【0041】

【発明の実施の形態】(第1の実施形態)図1は、本発明の第1の実施形態における音響モデル学習装置の構成を示す図である。以下、図1を用いて、本実施形態における音響モデル学習装置の構成について説明する。なお、本実施形態では、音響モデルとして連続混合確率分布によるHMMを用いる。上記の音響モデルでは、所定の音声からフレームごとに抽出された特微量を示す確率分布を用いることによって、上記のフレームごとに分割された音声の断片が状態として表現され、その状態が構

成単位となる。

【0042】音響モデル学習装置は、音声分析部101と、辞書部102と、フォワード・バックワード計算部103と、再評価部104と、ビタビ計算部105と、重み計算部106と、を有する。以下、図1を用いて音響モデル学習装置の各部位について説明する。

【0043】音声分析部101には、音響モデルの学習に用いられる音声情報である学習用音声が入力される。なお、上記の学習用音声は、ビタビ計算部105にも入力される。

【0044】音声分析部101は、入力された学習用音声を所定周期ごとに区切り、その区間を「フレーム」として、フレームごとに学習用音声の周波数分析を行う。上記の分析の結果抽出されたフレームごとの学習用音声の（音響的）特徴量は、フォワード・バックワード計算部103および再評価部104に入力される。なお、特徴量としては、音声のパワーを用いてもよいし、パワーの変化量、ケプストラム、あるいはケプストラム変化量等を用いてもよい。

【0045】辞書部102には、音響モデルおよび正解列が入力される。上記の正解列は、所定の入力手段（図示せず）により入力される文字列の情報としてもよい。所定の入力手段は、音声分析部101およびビタビ計算部105に入力された学習用音声の内容を示す文字情報を正解列として辞書部102に入力する。

【0046】また、辞書部102は、入力された音響モデル（以下、入力音響モデル）と入力された正解列とに基づいて、サブワードモデルによる学習用辞書を作成し、格納する。なお、サブワードモデルによる学習用辞書とは、入力された正解列（例えば、実際に存在する単語等）を、音素あるいは音節単位（サブワード単位）等に分割した状態列の情報である。また、辞書部102は、学習用辞書とは別に、任意の文字列の情報である「任意の文字列を表す辞書」を予め格納している。

【0047】フォワード・バックワード計算部103は、辞書部102に格納されている学習用辞書を参照し、音声分析部101により抽出された学習用音声の特徴量と、入力された入力音響モデルと、に基づいて、フォワード・バックワード法によるフォワード確率とバックワード確率とを算出する。さらに、フォワード・バックワード計算部103は、算出したフォワード確率とバックワード確率とに基づいて、学習用音声の特徴量と入力音響モデルの状態との間の対応確率を算出する。フォワード・バックワード計算部103は、算出した対応確率を再評価部104へ出力する。

【0048】フォワード・バックワード計算部103は、入力された学習用音声から変換されたフレーム t ごとの特徴量を O_t （ t は1以上T以下の整数）として、

フォワード確率 α を、以下に示す（式1.1）および（式1.2）に基づいて算出する。また、フォワード・バックワード計算部103は、バックワード確率 β を、（式2.1）および（式2.2）により示されている式に基づいて算出する。

【0049】また、フォワード・バックワード計算部103は、算出したフォワード確率 α とバックワード確率 β とを用いて、対応確率 γ を、（式3.1）により示される式に基づいて算出する。

【0050】ビタビ計算部105には、音声分析部101と同様の学習用音声が入力される。また、ビタビ計算部105には、辞書部102を介して入力音響モデルが入力される。

【0051】ビタビ計算部105は、入力された学習用音声を所定時間（フレーム）ごとに分割する。次に、ビタビ計算部105は、所定の文字情報を参照して、上記の各フレームに入力音響モデルに基づく状態あるいは複数の状態からなる状態列を最尤に割り当て、ビタビマッチング（Viterbi Matching）を行い、所定の最尤状態列を作成する。

【0052】重み計算部106は、ビタビ計算部105により複数種類の所定の文字情報を参照して作成された複数種類の最尤状態列に基づいて重み係数 R_t を算出する。

【0053】再評価部104は、重み計算部106により算出された重み係数 R_t と、フォワード・バックワード計算部103により算出された対応確率と、音声分析部101により抽出された特徴量と、フォワード・バックワード計算部103を介して入力された入力音響モデルと、に基づいて、音響モデルの各状態の統計量（混合重み、平均ベクトル、および共分散行列の各平均）を計算する。再評価部104は、抽出された統計量に基づいて、入力音響モデルの各パラメータ（混合重み、平均ベクトル、および共分散行列の各平均）を再評価する。再評価部104は、入力音響モデルの各パラメータの再評価に基づいて、音響モデルを作成する。再評価部104は、作成した音響モデルを、出力音響モデルとして出力する。

【0054】再評価部104は、対応確率 γ に重み係数 R_t を積算して重み付けを行う。再評価部104は、重み付けされた対応確率 $\gamma \cdot R_t$ を用いて、混合重み c_{jk} 、平均ベクトル $\mu(t, j, k)$ 、および共分散行列 $U(j, k)$ の各平均を統計量として算出する。上記の統計量は、以下に示す（式5.1）、（式5.2）、および（式5.3）により与えられる。

【0055】

【数5】

$$c_{jk} = \frac{\sum_{t=1}^T \gamma(t, j, k) \cdot R_t}{\sum_{t=1}^T \sum_{k=1}^K \gamma(t, j, k) \cdot R_t} \quad \dots \quad (\text{式5.1})$$

$$\bar{\mu}(t, j, k) = \frac{\sum_{t=1}^T \gamma(t, j, k) \cdot O_t \cdot R_t}{\sum_{t=1}^T \gamma(t, j, k) \cdot R_t} \quad \dots \quad (\text{式5.2})$$

$$\bar{U}(j, k) = \frac{\sum_{t=1}^T \gamma(t, j, k) \cdot (O_t - \mu_{jk}) (O_t - \mu_{jk})' \cdot R_t}{\sum_{t=1}^T \gamma(t, j, k) \cdot R_t} \quad \dots \quad (\text{式5.3})$$

【0056】なお、混合重み c_{jk} は、HMMにおける状態 S_j の k 番目の混合分布要素に対する混合重みである。また、平均ベクトル $\mu(t, j, k)$ は、HMMにおける状態 S_j の k 番目の混合分布要素の平均ベクトルである。また、共分散行列 $U(j, k)$ は、HMMにおける状態 S_j の k 番目の混合分布要素の共分散行列である。また、 V_k は、文字列 V における所定の文字を示す。また、 $(O_t - \mu_{jk})'$ は、ベクトル $(O_t - \mu_{jk})$ の対置ベクトルを表す。

【0057】図2は、本発明の第1の実施形態における入力音響モデルが表現可能な音素セットを示す図である。また、図3は、本発明の第1の実施形態における音響モデル学習装置が作成する学習用辞書を示す図である。また、図4は、本発明の第1の実施形態における重み係数 R_t を示す図である。また、図9は、本発明の第1の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。以下、図1～4を用い、図9に沿って本実施形態における音響モデル学習装置の動作について説明する。

【0058】本実施形態では、学習用音声の一例として、所定の話者による「加藤今太郎（かとうこんたろう）」の発声を用いる。また、本実施形態では、入力音響モデル（初期モデル）として、上記の所定の話者による「かとうこんたろう」の発声を、「さとうこんたろう」と認識する音響モデルが与えられたとする。

【0059】なお、HMMでは、1状態に対応する音声の長さは可変であり、ビタビマッチング等を用いることにより、HMMにおける最尤な状態系列が得られる。しかしながら、本実施形態では、簡単のために、入力音声

は14フレームの音声であり、1フレームにつき1状態が割り当てられているものとする。

【0060】まず、所定の制御手段（図示せず）は、学習用音声が音声分析部101に入力されたか否かを判断する（ステップS901）。学習用音声が音声分析部101に入力されていないと判断された場合（ステップS901/N_o）、ステップS901の工程が繰り返される。

【0061】学習用音声が音声分析部101に入力されたと判断された場合（ステップS901/Y_es）、音声分析部101は、フレームごとに学習用音声の周波数を分析し、その分析した学習用音声の周波数に基づいて学習用音声の特徴量を抽出する（ステップS902）。抽出した学習用音声の特徴量は、フォワード・バックワード計算部103および再評価部104へ出力される。

【0062】次に、所定の制御手段は、正解列および入力音響モデルが辞書部102に入力されたか否かを判断する（ステップS903）。正解列および入力音響モデルが入力されていないと判断された場合（ステップS903/N_o）、ステップS903の工程が繰り返される。

【0063】正解列および入力音響モデルが辞書部102に入力されたと判断された場合（ステップS903/Y_es）、辞書部102は、入力された正解列と入力音響モデルとに基づいて学習用辞書を作成し、作成した学習用辞書を格納する（ステップS904）。

【0064】ここで、図2および図3を用いて、辞書部102が学習用辞書を作成する工程について説明する。図2には、本実施形態における入力音響モデルが表現で

きる音素の列（音素セット）が示されている。上記の音素セットは、入力音響モデルに含まれている。辞書部102は、上記の音素セットを用いて、学習用音声「かとうこんたろう」を「k-a-t-o-u-k-o-n g-t-a-r-o-u」と音素単位に分割する。分割した音素を、状態 S_i （ i は1以上13以下の整数）にそれぞれ割り当て、図3に示されるような状態列、すなわち学習用音声に対応する学習用辞書を作成する。辞書部102は、作成した学習用辞書を格納する。

【0065】辞書部102による学習用辞書作成後、フォワード・バックワード計算部103は、辞書部102により作成された学習用辞書を参照し、音声分析部101により抽出された特徴量に基づいて、フォワード確率およびバックワード確率を算出する（ステップS905）。

【0066】次に、フォワード・バックワード計算部103は、算出したフォワード確率とバックワード確率に基づいて対応確率を算出する（ステップS906）。

【0067】所定の制御手段は、音声分析部101に入力された学習用音声と同様の学習用音声がビタビ計算部105に入力されたか否かを判断する。また、所定の制御手段は、入力音響モデルが辞書部102を介してビタビ計算部105に入力されたか否かを判断する（ステップS907）。学習用音声および入力音響モデルがビタビ計算部105に入力されていないと判断された場合（ステップS907/N_o）、ステップS907の工程

$$R_t = 0 \quad (\exists j \ s.t. \sum_k \gamma(t, j, k) \neq \sum_k \gamma(t, j, k)) \dots \quad (\text{式6.1})$$

$$R_t = 1 \quad (\forall j, \sum_k \gamma(t, j, k) = \sum_k \gamma(t, j, k)) \dots \quad (\text{式6.2})$$

【0072】話者による誤発声あるいは品質の低い音声を学習用音声として用いた場合、入力された正解列と入力音響モデルにより認識される学習用音声との間で差異が発生する可能性、つまり、入力された学習用音声による所定の言語単位（例えば、音素単位、音節単位等）の音声サンプルが音響モデルにより誤認識される可能性が高い。上記の誤認識された音声サンプルが出力音響モデルに大きく反映しないことによって、信頼性の高い出力音響モデルを得ることが可能となる。

【0073】重み計算部106は、第1の最尤状態列における各状態と、第2の最尤状態列における各状態と、をフレームごとに比較し、上記の（式6.1）および（式6.2）に基づいて重み係数 R_t を算出する。

【0074】（式6.1）は、所定のフレームにおいて、第1の最尤状態列と第2の最尤状態列との間に差異が発生した場合の重み係数 R_t を与える式であり、上記の場合、重み係数 R_t は「0」として算出される。

【0075】（式6.2）は、全てのフレームにおいて、第1の最尤状態列と第2の最尤状態列とが一致した

が繰り返される。

【0068】学習用音声および入力音響モデルがビタビ計算部105に入力されたと判断された場合（ステップS907/Y_es）、ビタビ計算部105は、入力された学習用音声および入力音響モデルを用い、辞書部102により作成された学習用辞書を参照して、ビタビマッチングにより最尤状態列を生成する（ステップS908）。なお、学習用辞書を参照して生成された上記の最尤状態列を第1の最尤状態列とする。

【0069】さらに、ビタビ計算部105は、入力された学習用音声および入力音響モデルを用い、辞書部102に格納されている任意の文字列を表す辞書を参照して、ビタビマッチングにより最尤状態列を生成する（ステップS909）。なお、任意の文字を表す辞書を参照して生成された上記の最尤状態列を第2の最尤状態列とする。

【0070】次に、重み計算部106は、ビタビ計算部105により生成された第1の最尤状態列の各状態と第2の最尤状態列の各状態を比較し、以下に示す（式6.1）および（式6.2）により与えられる重み係数 R_t を算出する（ステップS910）。なお、重み係数 R_t は、学習用音声の各フレームにそれぞれ対応するように算出される。

【0071】

【数6】

場合の重み係数 R_t を与える式であり、上記の場合、重み係数 R_t は「1」として算出される。

【0076】話者の誤発声等により学習用音声の品質が低下した場合、その品質低下が生じた部分に対応するフレームに割り当てられている第1の最尤状態列の状態と、第2の最尤状態列の状態との間に差異が発生する。従って、信頼性の高い出力音響モデルを得るためにには、上記の差異が生じた部分が出力音響モデルに反映されないようにする必要がある。

【0077】本実施形態では、学習用音声における高品質部分（所定のフレームにおいて第1の最尤状態列の状態と第2の最尤状態列とが一致した状態）の重み係数 R_t を「1」とし、低品質部分の重み係数 R_t を高品質部分の重み係数 R_t よりも低い値である「0」とすることによって、学習用音声の低品質部分、すなわち学習用音声が入力音響モデルにより誤認識されている部分が出力音響モデルに反映されないようにしている。

【0078】本実施形態における入力音響モデルでは、学習用音声「かとうこんたろう」は、「さとうこんたろ

う」と認識される。上記のような場合、「か」の部分が実際にどのような発声であったか不明であるが、「か」の部分における音素「k」が入力音響モデルにより誤認識されている。音素「k」のモデルが正しく認識される出力音響モデルを作成するためには、「か」の部分の音素「k」が出力音響モデルに反映しないように設定される必要がある。

【0079】図14は、(式6.1)および(式6.2)により図3の学習用辞書に与えられる重み係数 R_t を示す図である。 R_t ($t = 1 \sim 13$) は、それぞれ S_i ($i = 1 \sim 13$) における重み係数である。図14に示されているように、「か」の部分の音素「k」 ($= S_1$) における重み係数 R_1 を「0」とし、他の音素 ($S_2 \sim S_{13}$) における重み係数 $R_2 \sim R_{13}$ を「1」とすることによって、「か」の部分の音素「k」が出力音響モデルに反映しないようにすることが可能となり、信頼性の高い音響モデルを作成することが可能となる。

【0080】なお、本実施形態では、重み係数 R_1 を「0」とすることにより、「か」の部分の音素「k」が出力音響モデルに反映しないようにしたが、重み係数 R_1 を「0以上1未満の任意の値」に設定することによって、「か」の部分の音素「k」が出力音響モデルに与える影響を調整することが可能となる。

【0081】以下、再び図9のフローチャートに沿って音響モデル学習装置の動作について説明を進める。再評価部104は、重み計算部106により算出された重み係数 R_t と、音声分析部101により抽出された特徴量と、フォワード・バックワード計算部103により算出された対応確率と、に基づいて、音響モデルの各統計量（混合重み、平均ベクトル、および共分散行列の各平均）を算出する（ステップS911）。

【0082】再評価部104は、音響モデルの各統計量算出後、算出した統計量に基づいて、フォワード・バックワード計算部103を介して入力された入力音響モデルの各パラメータ（混合重み分布、平均ベクトル、および共分散行列の各平均）を再評価し、出力音響モデルを作成する（ステップS912）。作成された出力音響モデルは、再評価部104から出力される（ステップS913）。出力音響モデル出力後、音響モデル学習装置は、動作を終了する。

【0083】（第2の実施形態）以下、特記しない限り、本発明の第2の実施形態における音響モデル学習装置の構成および動作は、本発明の第1の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【0084】一般に、騒音環境が学習用音声の品質を低下させる場合、学習用音声の誤認識は、单一の音素にとどまらず、その音素の周辺音素にも影響を与える。第1の実施形態では、重み係数 R_t を音素ごとに設定していたが、環境騒音などの理由により複数の音素にわたって

誤認識される場合、音節単位で重み付けを行うことによって、より信頼性の高い出力音響モデルを作成することが可能となる。

【0085】図5は、本発明の第2の実施形態における重み係数 R_t を示す図である。第1の実施形態と同様に重み係数 R_t ($t = 1 \sim 13$) は、それぞれ図3における状態 S_i ($i = 1 \sim 13$) に対応する。

【0086】第1の実施形態では、「か」の音素「k」 ($= S_1$) の重み係数 R_1 を「0」に設定していた。本実施形態では、学習用音声「かとうこんたろう」における音節「か (k-a)」において、品質が低下し、第1の最尤状態列と第2の最尤状態列との間に差異が生じている。上記のように音節単位で学習用音声の品質低下が生じている場合、「か」の音素「k」 ($= S_1$) の重み係数 R_1 と、音素「a」 ($= S_2$) の重み係数 R_2 と、をそれぞれ「0」に設定することによって、音素「k」 ($= S_1$) の重み係数 R_1 のみを「0」とする場合と比較して、より信頼性の高い出力音響モデルを作成することが可能となる。

【0087】なお、本実施形態では、重み係数 R_1 および R_2 を「0」とすることにより、「か」の部分の音素「k」および音素「a」が出力音響モデルに反映しないようにしたが、重み係数 R_1 および R_2 を「0以上1未満の任意の値」に設定することによって、「か」の部分の音素「k」および音素「a」が出力音響モデルに与える影響を調整することが可能となる。

【0088】（第3の実施形態）以下、特記しない限り、本発明の第3の実施形態における音響モデル学習装置の構成および動作は、本発明の第1の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【0089】第2の実施形態では、騒音環境による学習用音声の誤認識は、单一の音素にとどまらず、その音素の周辺音素にも影響を与える場合について説明した。第2の実施形態では、重み係数 R_t を音節ごとに設定していたが、環境騒音などの理由により誤認識される音素の範囲が音節単位よりもさらに広い範囲にわたって存在する場合、重み付けする音素の範囲を音節単位よりもさらに拡大し、単語単位とすることによって、より信頼性の高い出力音響モデルを作成することが可能となる。

【0090】図6は、本発明の第3の実施形態における重み係数 R_t を示す図である。第1の実施形態と同様に重み係数 R_t ($t = 1 \sim 13$) は、それぞれ図3における状態 S_i ($i = 1 \sim 13$) に対応する。

【0091】第1の実施形態では、「か」の音素「k」 ($= S_1$) の重み係数 R_1 を「0」に設定していた。また、第2の実施形態では、「か」の音素「k」 ($= S_1$) の重み係数 R_1 と、音素「a」 ($= S_2$) の重み係数 R_2 と、をそれぞれ「0」に設定していた。本実施形態では、学習用音声「かとうこんたろう」における单

語「かとう（k-a-t-o-u）」において、品質が低下し、第1の最尤状態列と第2の最尤状態列との間に差異が生じている。上記のように単語単位で学習用音声の品質低下が生じている場合、単語「かとう（k-a-t-o-u）」における音素「k」（=S₁）、音素「a」（=S₂）、音素「t」（=S₃）、音素「o」（=S₄）、および音素「u」（=S₅）それに対応する重み係数R₁～R₅を「0」とすることによって、音素単位あるいは音節単位で重み係数R_tを「0」とする場合と比較して、より信頼性の高い出力音響モデルを作成することが可能となる。

【0092】なお、本実施形態では、重み係数R₁～R₅を「0」とすることにより、「かとう」の部分の音素「k-a-t-o-u」が出力音響モデルに反映しないようにしたが、重み係数R₁～R₅を0以上1未満の任意の値に設定することによって、「かとう」の部分の音素「k-a-t-o-u」が出力音響モデルに与える影響を調整することが可能となる。

【0093】(第4の実施形態)以下、特記しない限り、本発明の第4の実施形態における音響モデル学習装

$$R_t = 10 \quad (\exists j, s.t. \sum_k \gamma(t, j, k) \neq \sum_k \gamma(t, j, k)) \dots \quad (式7.1)$$

$$R_t = 1 \quad (\forall j, \sum_k \gamma(t, j, k) = \sum_k \gamma(t, j, k)) \dots \quad (式7.2)$$

【0097】本実施形態では、第1の実施形態と同様に、所定の話者により入力された「かとうこんたろう」という学習用音声を、「さとうこんたろう」と認識する音響モデルが入力される。第1の実施形態では、「か」の音素「k」（=S₁）に対応する重み係数R₁を「0」に設定し、出力音響モデルに反映しないようにして、信頼性の高い出力音響モデルを作成していた。本実施形態では、第1の最尤状態列と第2の最尤状態列との間で差異が発生した「か」の音素「k」（=S₁）に、第1の最尤状態列と第2の最尤状態列との間で一致した他の音素に設定された「重み係数R_t=1（t=2～13）」よりも高い「重み係数R₁=10」を設定する。

【0098】上記のように、「重み係数R₁=10」と設定することによって、十分に学習されていない稀な特徴と考えられる「か」の音素「k」（=S₁）を、他の音素よりも出力音響モデルに大きく反映させることが可能となる。

【0099】なお、本実施形態では、重み係数R_tによる重み付けを音素単位で行ったが、第2の実施形態のように音節単位で行ってもよいし、第3の実施形態のように単語単位で行ってもよい。

【0100】また、本実施形態では、正解列と入力音響モデルにより認識された学習用音声との間で差異が生じた音素に対応する重み係数R_tを「10」としたが、正

置の構成および動作は、本発明の第1の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【0094】上記の第1から第3の実施形態では、第1の最尤状態列と第2の最尤状態列との間で差異が生じた部分（学習用音声の品質が低下した部分）の重み係数R_tを「0」に設定し、出力音響モデルに反映されないようにしていた。本実施形態における音響モデル学習装置は、学習用音声における誤発声あるいは品質の低い音声が生じた部分を発声の一変化として積極的に取り入れ、学習用音声の高品質部分よりも高い重み係数R_tを設定することによって、低品質の学習用音声のサンプル数を増加させ、低品質の学習用音声に対する認識性能を向上させる。

【0095】図7は、本発明の第4の実施形態における重み係数R_tを示す図である。図7に示される重み係数R_tは、以下に示す（式7.1）および（式7.2）により与えられる。

【0096】

【数7】

$$R_t = 10 \quad (\exists j, s.t. \sum_k \gamma(t, j, k) \neq \sum_k \gamma(t, j, k)) \dots \quad (式7.1)$$

$$R_t = 1 \quad (\forall j, \sum_k \gamma(t, j, k) = \sum_k \gamma(t, j, k)) \dots \quad (式7.2)$$

解列と学習用音声との間で一致した音素と比較して大きな数値であれば、差異が生じた音素に対応する重み係数R_tは、他の値であってもよい。

【0101】(第5の実施形態)以下、特記しない限り、本発明の第5の実施形態における音響モデル学習装置の構成および動作は、本発明の第1の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【0102】統計モデルの信頼性は、統計モデルのパラメータ学習に用いられた音声サンプル（音素、音節、あるいは単語）の量により大きく影響される。従って、各音響モデルにおける信頼性を均一化するためには、入力される各音声サンプルの量に著しい偏りが生じないようになる必要がある。

【0103】本実施形態では、第1の最尤状態列における各状態ごとの重み係数R_tの和を一定にし、入力される所定の言語単位（音素、音節、あるいは単語等）の各音声サンプルにおけるサンプル数を均一化する。

【0104】図10は、本発明の第5の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。以下、図10を用い、図10に沿って、本実施形態における音響モデル学習装置の動作について説明する。

【0105】本実施形態では、第1の実施形態と同様に、学習用音声の一例として、所定の話者による「加藤

今太郎（かとうこんたろう）」の発声を用いる。

【0106】まず、所定の制御手段（図示せず）は、学習用音声が音声分析部101に入力されたか否かを判断する（ステップS1001）。学習用音声が音声分析部101に入力されていないと判断された場合（ステップS1001/N_o）、ステップS1001の工程が繰り返される。

【0107】学習用音声が音声分析部101に入力されたと判断された場合（ステップS1001/Y_es）、音声分析部101は、フレームごとに学習用音声の周波数を分析し、その分析した学習用音声の周波数に基づいて学習用音声の特徴量を抽出する（ステップS1002）。抽出した学習用音声の特徴量は、フォワード・バックワード計算部103および再評価部104へ出力される。

【0108】次に、所定の制御手段は、正解列および入力音響モデルが辞書部102に入力されたか否かを判断する（ステップS1003）。正解列および入力音響モデルが入力されていないと判断された場合（ステップS1003/N_o）、ステップS1003の工程が繰り返される。

【0109】正解列および入力音響モデルが辞書部102に入力されたと判断された場合（ステップS1003/Y_es）、辞書部102は、入力された正解列と入力音響モデルとに基づいて学習用辞書を作成し、作成した学習用辞書を格納する（ステップS1004）。

【0110】辞書部102による学習用辞書作成後、フォワード・バックワード計算部103は、辞書部102により作成された学習用辞書を参照し、音声分析部101により抽出された特徴量に基づいて、フォワード確率およびバックワード確率を算出する（ステップS1005）。

$$R_t = 1 \quad (\sum_{\substack{n \leq t \\ 0_t \leq v_n}} R_n \leq M)$$

$$R_t = 0 \quad (\sum_{\substack{n \leq t \\ 0_t \leq v_n}} R_n > M)$$

$$s.t. \quad M = \min_j \sum_k \gamma(t, j, k) \quad \dots \quad (式9.3)$$

【0117】本実施形態では、上記の（式8. 1）で与えられる条件により、学習用音声を構成する同一の音声サンプル（音素、音節、あるいは単語単位）が割り当てられている状態ごとに重み係数R_tの和をとり、重み係数R_tの和が等しくなるように、重み係数R_tを算出することによって、各音声サンプルがそれぞれ出力音響モデルに与える影響が均一になる。

【0118】本実施形態では、本発明の第1の実施形態

【0111】次に、フォワード・バックワード計算部103は、算出したフォワード確率とバックワード確率とに基づいて対応確率を算出する（ステップS1006）。

【0112】所定の制御手段は、音声分析部101に入力された学習用音声と同様の学習用音声がビタビ計算部105に入力されたか否かを判断する。また、所定の制御手段は、入力音響モデルが辞書部102を介してビタビ計算部105に入力されたか否かを判断する（ステップS1007）。学習用音声および入力音響モデルがビタビ計算部105に入力されていないと判断された場合（ステップS1007/N_o）、ステップS1007の工程が繰り返される。

【0113】学習用音声および入力音響モデルがビタビ計算部105に入力されたと判断された場合（ステップS1007/Y_es）、ビタビ計算部105は、入力された学習用音声および入力音響モデルを用い、辞書部102により作成された学習用辞書を参照して、ビタビマッチングにより最尤状態列を生成する（ステップS1008）。なお、学習用辞書を参照して生成された上記の最尤状態列を第1の最尤状態列とする。

【0114】次に、重み計算部106は、ビタビ計算部105により生成された第1の最尤状態列の各状態を参照し、以下に示す（式8. 1）、（式9. 1）、（式9. 2）、および（式9. 3）に基づいて、重み係数R_tを算出する（ステップS1009）。

【0115】

【数8】

$$\forall i, j, \sum_{\substack{0_n=v_i \\ 0_n=v_j}} R_n = \sum_{\substack{0_n=v_i \\ 0_n=v_j}} R_n \quad \dots \quad (式8.1)$$

【0116】

【数9】

$$\dots \quad (式9.1)$$

$$\dots \quad (式9.2)$$

と同様に図3に示される学習用辞書が生成されるとする。図8は、本発明の第5の実施形態における重み係数R_tを示す図である。図8に示される重み係数R_tは、上記の（式9. 1）、（式9. 2）および（式9. 3）に基づいて設定されている。なお、図8における重み係数R_t（t=1～13）は、図3に示されている状態S_i（i=1～13）にそれぞれ対応している。

【0119】本実施形態では、割り当てられたフレーム

の値が小さなものから順に、学習用音声を構成する音素を観測した場合、初めて観測された種類の音素に対応する重み係数 R_t を「1」とし、以前観測された種類の音素に対応する重み係数 R_t を「0」としている。

【0120】以下、図3および図8を用いて説明すると、例えば、 S_6 の音素「k」は、すでに S_1 において観測されているので重み係数 R_6 は「0」に設定されている。一方、 S_{11} の音素「r」は、 $S_1 \sim S_{10}$ において観測されていないので重み係数 R_{11} は「1」に設定されている。

【0121】上記のように重み係数 R_t が算出されることによって、同一種類の音素に付加されている重み係数 R_t の和は、それぞれ「1」となり、各音素が音声サンプルとして収集される回数が均等となる。

【0122】以下、再び図10のフローチャートに沿って音響モデル学習装置の動作について説明を進める。再評価部104は、重み計算部106により算出された重み係数 R_t と、音声分析部101により抽出された特徴量と、フォワード・バックワード計算部103により算出された対応確率と、に基づいて、音響モデルの各統計量（混合重み、平均ベクトル、および共分散行列の各平均）を算出する（ステップS1010）。

【0123】再評価部104は、音響モデルの各統計量算出後、算出した統計量に基づいて、フォワード・バックワード計算部103を介して入力された入力音響モデルの各パラメータ（混合重み分布、平均ベクトル、および共分散行列の各平均）を再評価し、出力音響モデルを作成する（ステップS1011）。作成された出力音響モデルは、再評価部104から出力される（ステップS1012）。出力音響モデル出力後、音響モデル学習装置は、動作を終了する。

【0124】本実施形態では、以上説明したように、同一の音声サンプル（音素、音節、あるいは単語）が割り当てられた状態ごとの重み係数 R_t の和を一定とすることによって、各音声サンプル（音素、音節、あるいは単語単位）のサンプル量および出力音響モデルに与える影響を均一化し、信頼性の高い出力音響モデルを作成することを可能としている。

【0125】また、音響モデル学習装置は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析処理と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとの特徴量を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成処理と、辞書生成処理により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出処理と、所定

の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成処理と、最尤状態列生成処理により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算処理と、対応確率算出処理により算出された対応確率と、重み計算処理により算出された重み係数と、音声分析処理により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価処理と、を行う。上記の処理は、音響モデル学習装置が有するコンピュータプログラムにより実行されるが、上記のプログラムは、光ディスクあるいは磁気ディスク等の記録媒体に記録され、上記の記録媒体からロードされるようにしてもよい。

【0126】なお、上記の実施形態は本発明の好適な実施の一例であり、本発明の実施形態は、これに限定されるものではなく、本発明の要旨を逸脱しない範囲において種々変形して実施することが可能となる。

【0127】

【発明の効果】以上説明したように、本発明は、学習用音声のフレームごとに重み係数を算出し、上記の重み係数による重み付けを出力音響モデルに反映させることによって、観測された音声サンプルのうち音響モデルの作成に有用なだけを抽出し、信頼性の高い音響モデルを作成することが可能となる。

【0128】また、本発明は、品質の高い所定の言語単位（音素、音節、あるいは単語等）の音声サンプルの重み付け係数を「1」とし、品質の低い音声サンプルの重み付け係数を「0」とすることによって、品質の低い音声サンプルが出力音響モデルに反映しないようにすることが可能となる。

【0129】また、本発明は、品質の高い所定の言語単位の音声サンプルの重み付け係数を「1」とし、品質の低い音声サンプルの重み付け係数を「1より大きな任意の値」とすることによって、品質の低い音声サンプルに対する音声認識の精度が高い出力音響モデルを作成することが可能となる。

【0130】また、本発明は、同一の音声サンプル（音素、音節、あるいは単語）が割り当てられた状態ごとの重み係数の和を一定とすることによって、各音声サンプル（音素、音節、あるいは単語単位）のサンプル量および出力音響モデルに与える影響を均一化し、信頼性の高い出力音響モデルを作成することが可能となる。

【図面の簡単な説明】

【図1】本発明の第1の実施形態における音響モデル学習装置の構成を示す図である。

【図2】本発明の第1の実施形態における入力音響モ

ルが表現可能な音素セットを示す図である。

【図3】本発明の第1の実施形態における音響モデル学習装置が作成する学習用辞書を示す図である。

【図4】本発明の第1の実施形態における重み係数R_tを示す図である。

【図5】本発明の第2の実施形態における重み係数R_tを示す図である。

【図6】本発明の第3の実施形態における重み係数R_tを示す図である。

【図7】本発明の第4の実施形態における重み係数R_tを示す図である。

【図8】本発明の第5の実施形態における重み係数R_tを示す図である。

を示す図である。

【図9】本発明の第1の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。

【図10】本発明の第5の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。

【符号の説明】

101 音声分析部

102 辞書部

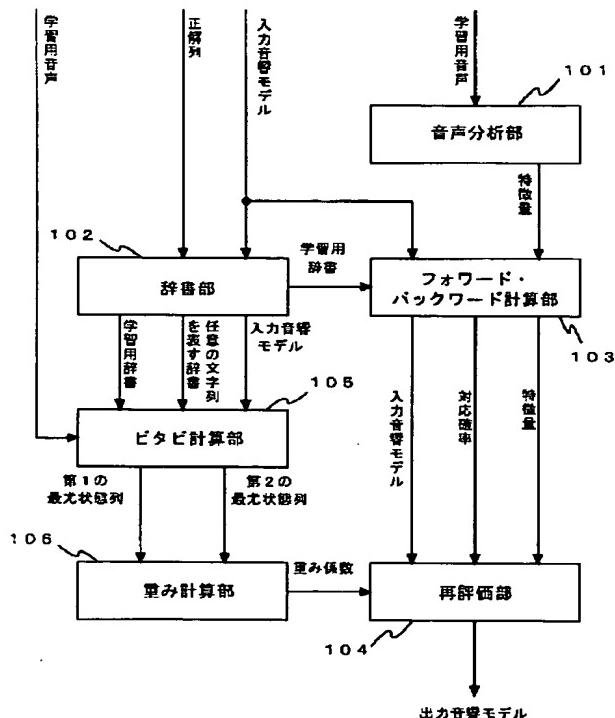
103 フォワード・バックワード計算部

104 再評価部

105 ビタビ計算部

106 重み計算部

【図1】



【図4】

R 1 = 0
R 2 = 1. 0
R 3 = 1. 0
R 4 = 1. 0
R 5 = 1. 0
R 6 = 1. 0
R 7 = 1. 0
R 8 = 1. 0
R 9 = 1. 0
R 10 = 1. 0
R 11 = 1. 0
R 12 = 1. 0
R 13 = 1. 0

【図8】

R 1 = 1. 0
R 2 = 1. 0
R 3 = 1. 0
R 4 = 1. 0
R 5 = 1. 0
R 6 = 0
R 7 = 0
R 8 = 1. 0
R 9 = 0
R 10 = 0
R 11 = 1. 0
R 12 = 0
R 13 = 0

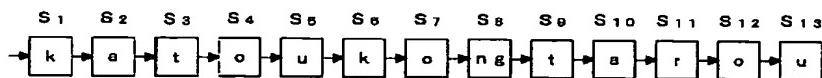
【図5】

R 1 = 0
R 2 = 0
R 3 = 1. 0
R 4 = 1. 0
R 5 = 1. 0
R 6 = 1. 0
R 7 = 1. 0
R 8 = 1. 0
R 9 = 1. 0
R 10 = 1. 0
R 11 = 1. 0
R 12 = 1. 0
R 13 = 1. 0

【図2】

a, b, ch, d, e, f, g, h, i, j, k, m, n, ng, o, p, r, s, sh, t, ts, u, v, w, y, z, zh

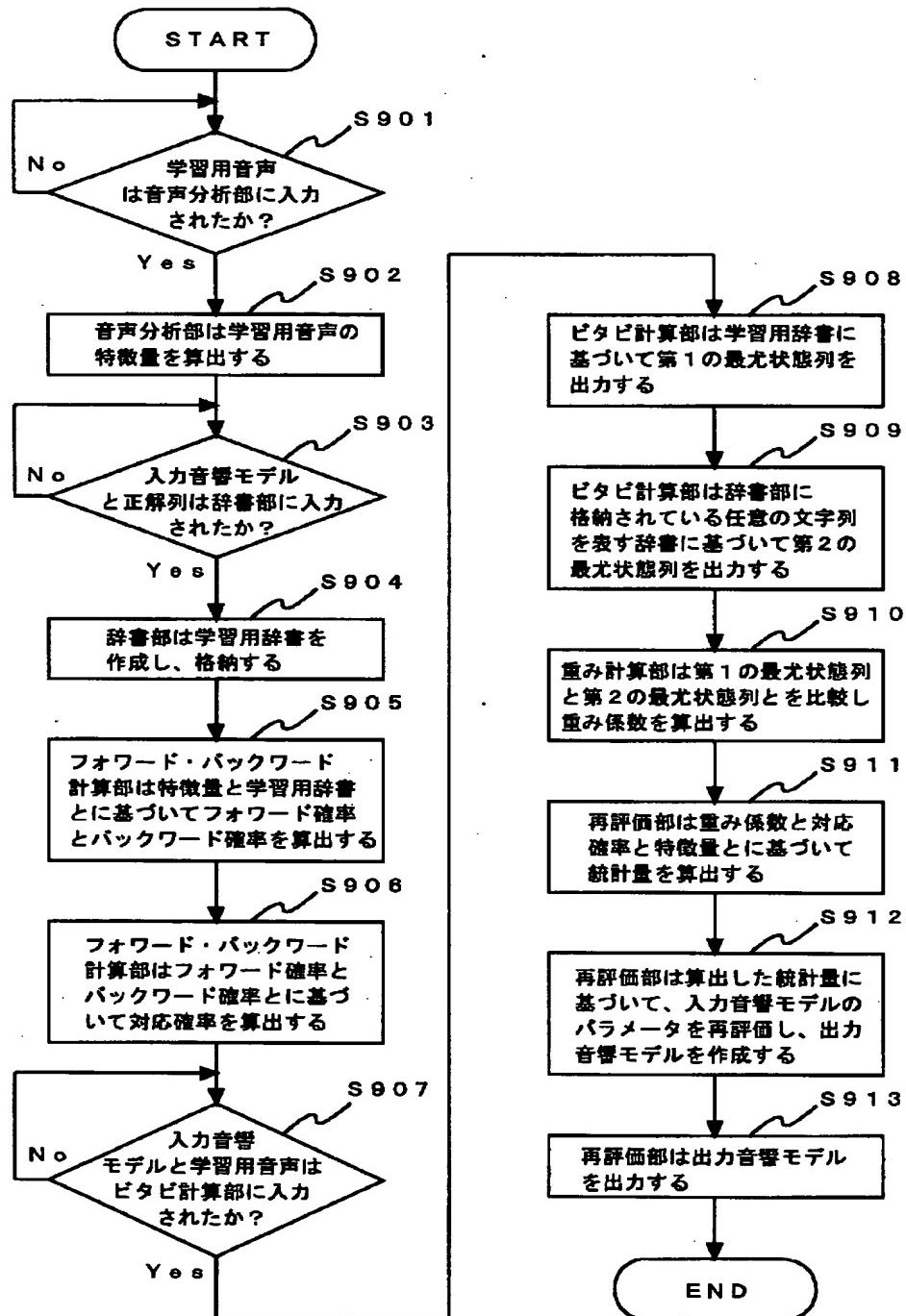
【図3】



【図7】

R 1 = 1. 0
R 2 = 1. 0
R 3 = 1. 0
R 4 = 1. 0
R 5 = 1. 0
R 6 = 1. 0
R 7 = 1. 0
R 8 = 1. 0
R 9 = 1. 0
R 10 = 1. 0
R 11 = 1. 0
R 12 = 1. 0
R 13 = 1. 0

【図9】



【図10】

